- Remember errer ε_i's are independent but the residuals e_i's are not since we have n observations but the degrees of freedom is only $n-k$!

- e_i's are the observed values of ε_i

- It is common to think of the residuals as the values of the errors. Hence, we can test the underlaying assumptions on the residuals.

One idea is to plot the residuals in an effective way to test how well the model fits the data & how well the assumptions are satisfied.

* Types of Residuals :

1) Regular Residuals (e_i) :

$$e_i = Y_i - \hat{Y}_i$$

$$\begin{aligned} e &= Y - \hat{Y} \\ &= Y - X\hat{\beta} \\ &= Y - X(X^TX)^{-1}X^TY \\ &= (I - X(X^TX)^{-1}X^T)Y \\ &= (I - H)Y \end{aligned}$$

H is symmetric and idempotent as :

$$H^2 = (X(X^TX)^{-1}X^T)(X(X^TX)^{-1}X^T)^T$$

$$
\begin{aligned}
&= (X(X^TX)^{-1}X^T)(X(X^TX)^{-1}X^T) \qquad \text{switch the order!}\\
&= X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^T\\
&= X(X^TX)^{-1}X^T\\
&= H
\end{aligned}
$$

H is called the Hat Matrix.

$$
\begin{aligned}
\text{Hence, } e &= (I-H)Y\\
&= (I-H)(X\beta+\varepsilon)\\
&= X\beta - HX\beta + (I-H)\varepsilon\\
&= X\beta - X(X^TX)^{-1}X^TX\beta + (I-H)\varepsilon\\
&= X\beta - X\beta + (I-H)\varepsilon\\
&= (I-H)\varepsilon
\end{aligned}
$$

$$
\begin{aligned}
Var(e) &= Var((I-H)\varepsilon)\\
&= (I-H)^2\, Var(\varepsilon)\\
&= (I-2H+H^2)\sigma^2\\
&= (I-2H+H)\sigma^2\\
&= (I-H)\sigma^2
\end{aligned}
$$

$$
= \begin{pmatrix}
Var(e_1) & Cov(e_1,e_2) & \cdots & Cov(e_1,e_n)\\
Cov(e_2,e_1) & Var(e_2) & \cdots & Cov(e_2,e_n)\\
 & & \ddots &
\end{pmatrix}
$$

$$
Var(e_i) = (1-h_{ii})\sigma^2 \quad , \quad h_{ii} = x_i(X^TX)^{-1}x_i
$$

$$
Cov(e_i,e_j) = (0-h_{ij})\sigma^2 = -h_{ij}\sigma^2 \quad , \quad I_{ij} = 0 \text{ (diagonal)}
$$

h_{ii} measures the distance of the i^{th} observation from the center of the x coordinate $\quad 0 \le h_{ii} \le 1$.

b) Studentized Residuals (r_i) :

$$Var(e_i) = \sigma^2 (1 - h_{ii}) = MS_{Res} (1 - h_{ii})$$
$$E(e_i) = 0$$

$$r_i = \frac{e_i}{\sqrt{MS_{Res}(1 - h_{ii})}}$$

studentized residuals have a constant variance $Var(r_i) = 1$ regardless of the location of the x-coordinate when the form of the model is correct.

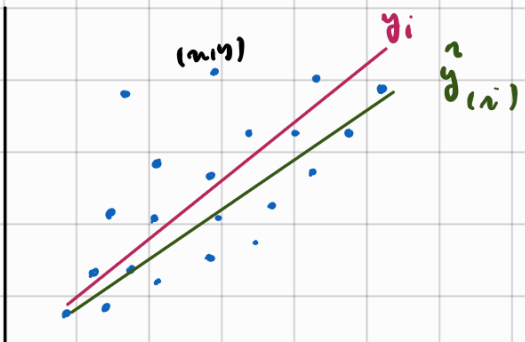
c) Standardized Residuals (d_i) :

$$Var(e_i) \approx \sigma^2$$

$$d_i = \frac{e_i}{\sqrt{MS_{Res}}}$$

d) Press Residuals ($e_{(i)}$)

i^{th} press residual : $e_{(i)} = y_i - \hat{y}_{(i)}$, $e_i \neq e_{(i)}$



- predicted line with all the observations
- predicted line with the

observation (x, y)

This residual is useful in the presence of an outlier

$\hat{y}_i$ is the fitted value of the i^{th} response based on all the observed except the i^{th}.

e.g : $M_1 \rightarrow Y = 2X + 3$ $\qquad M = (2, 4)$

$M_2 \rightarrow Y = 5X + 9$

$\hat{Y}_{M_1} = 7 \qquad Y - \hat{Y}_{M_1} = 4 - 7 = -3$

$\hat{Y}_{M_2} = 19 \qquad Y - \hat{Y}_{M_2} = 4 - 19 = -15$ ++

- We delete the i^{th} observation then fit the response model to the remaining $(n-1)$ observations and predict y_i.

- $e_{(i)} = y_i - \hat{y}_{(i)} = \dfrac{e_i}{1 - h_{ii}}$

- If we have very large press residual, we can identify the observations where the model does not fit the data well.

- Press Statistic $= \sum_{i=1}^{n} e_{(i)}^2$

$$= \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2$$

$$= \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2$$

which measures how well a regression model will perform in predicting a new data point.

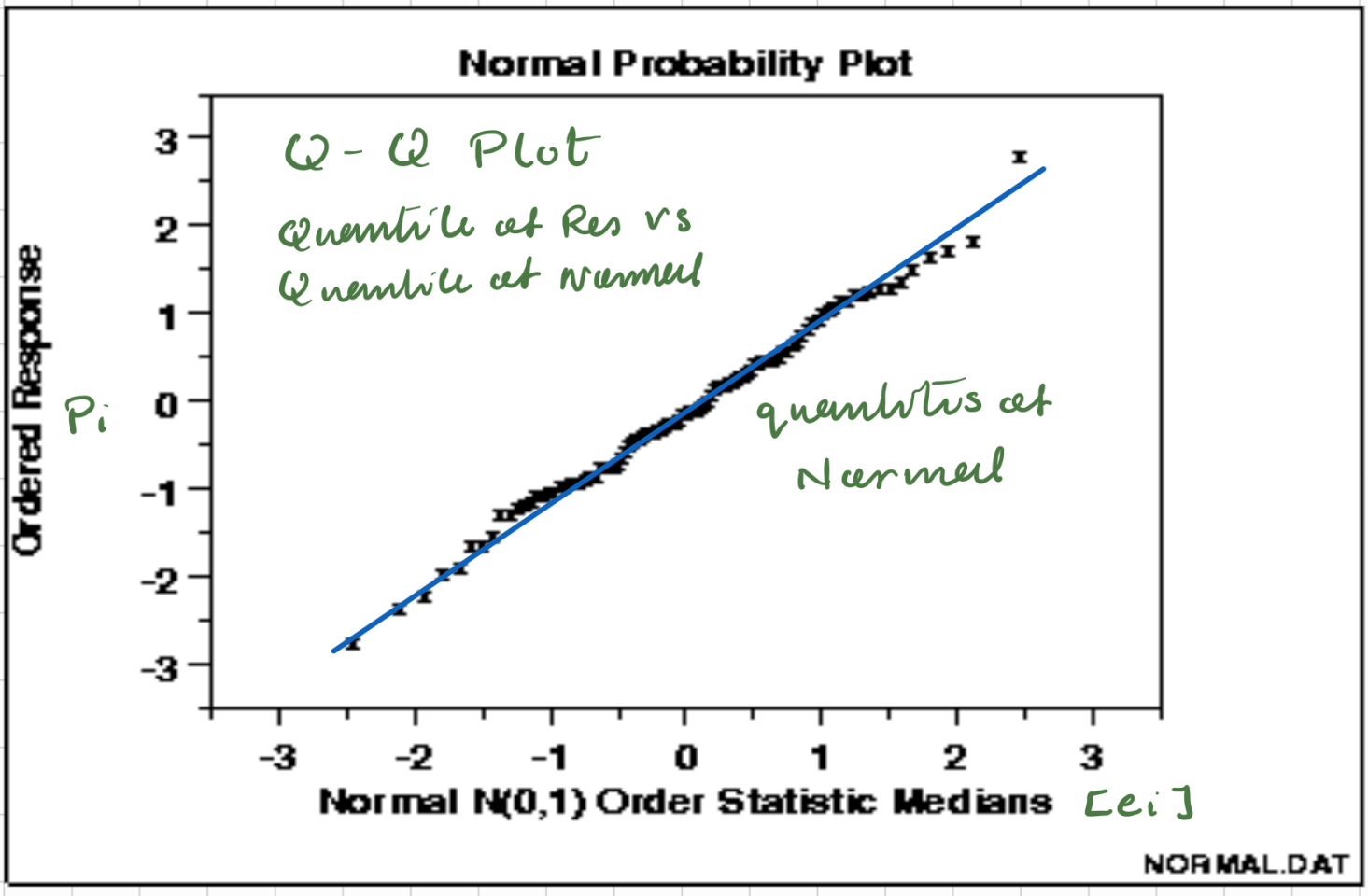
* Types of Residual Plots :

1) Normality Plot :

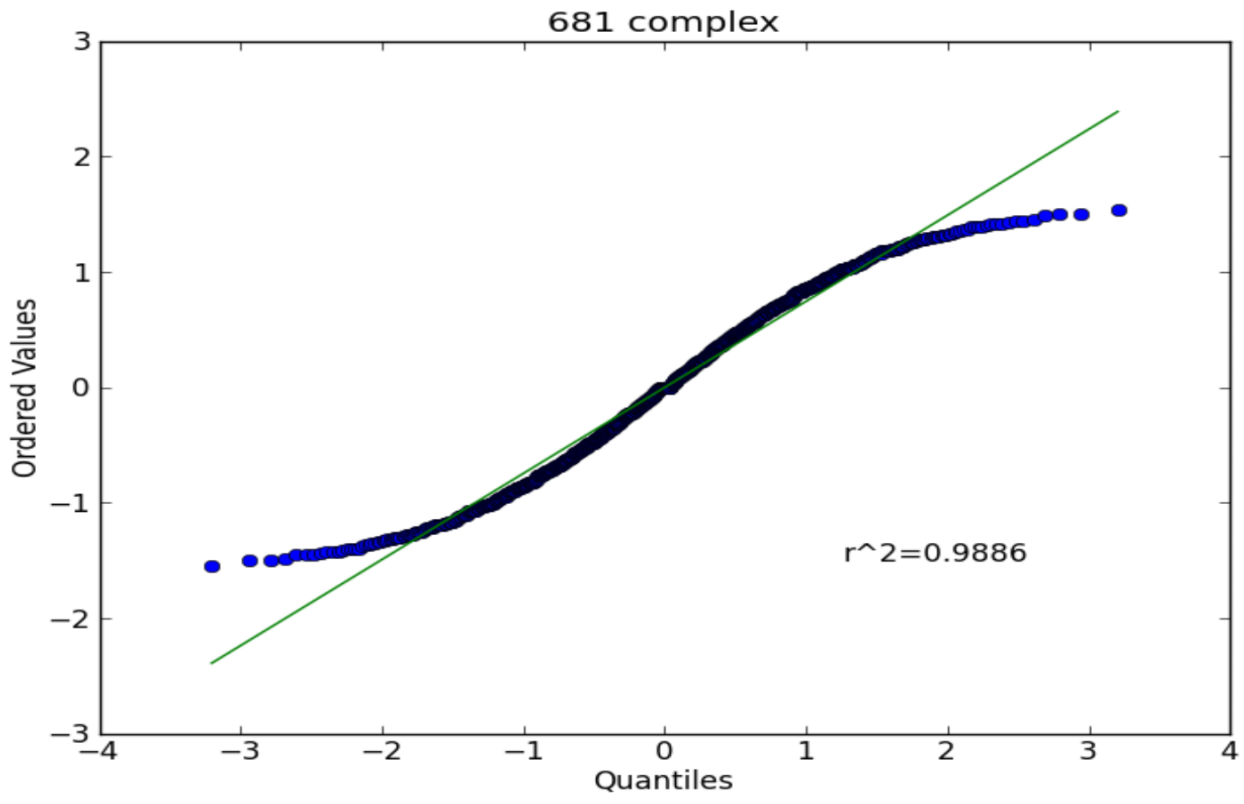
Let $e_1, \ldots, e_n$ be the n residuals, $e_{[1]}, \ldots, e_{[n]}$ are the residuals in an increasing order of magnitude.

Plot $[e_i]$ against Cumulative Probabilities :

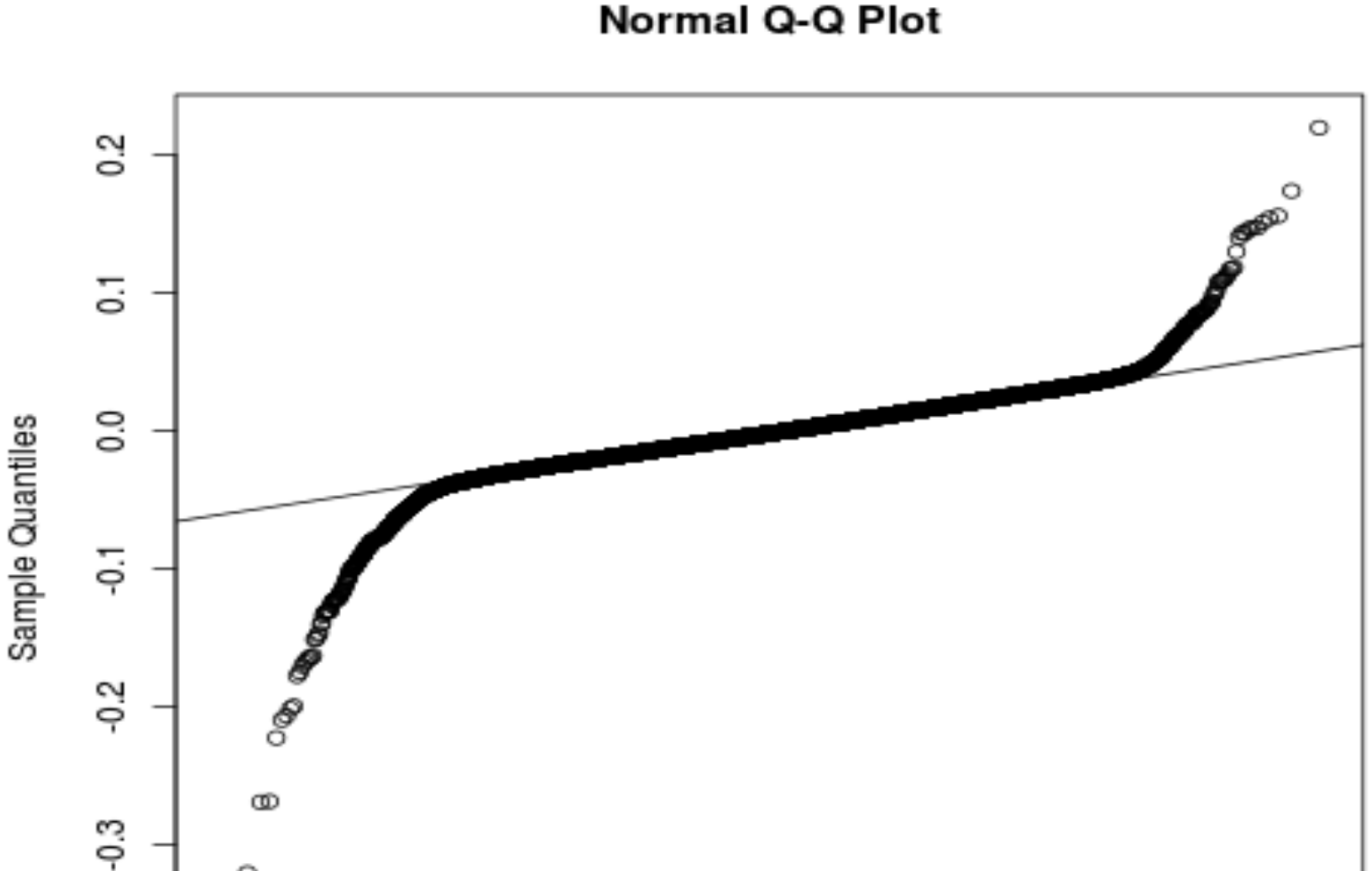
$$P_i = \frac{i - 1/2}{n}$$

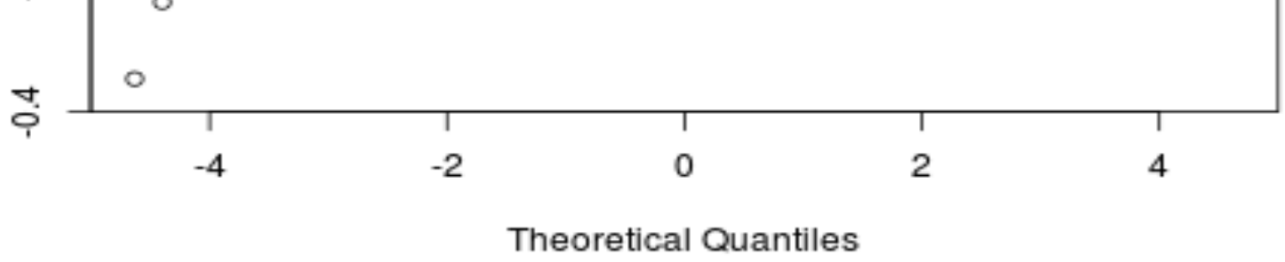


Normal Distribution (Ideal Situation)



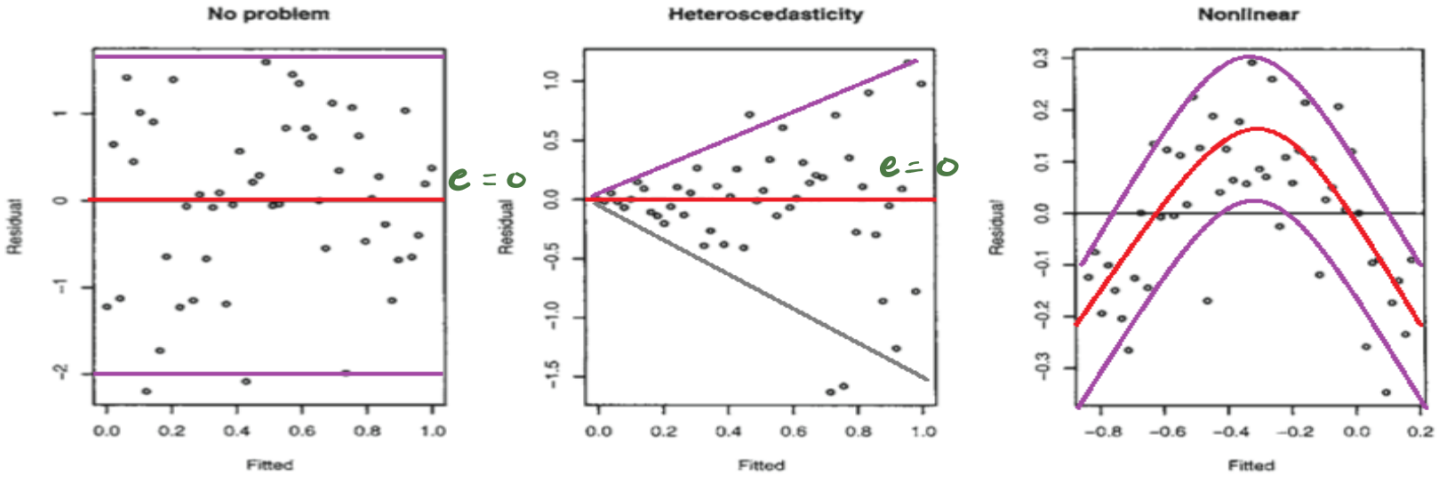
Light Tailed Distribution





Heavy Tailed Distribution

2) Plot of Residuals e_i vs fitted values ($\hat{y}_i$).



Satisfactory Model

Unsatisfactory Model

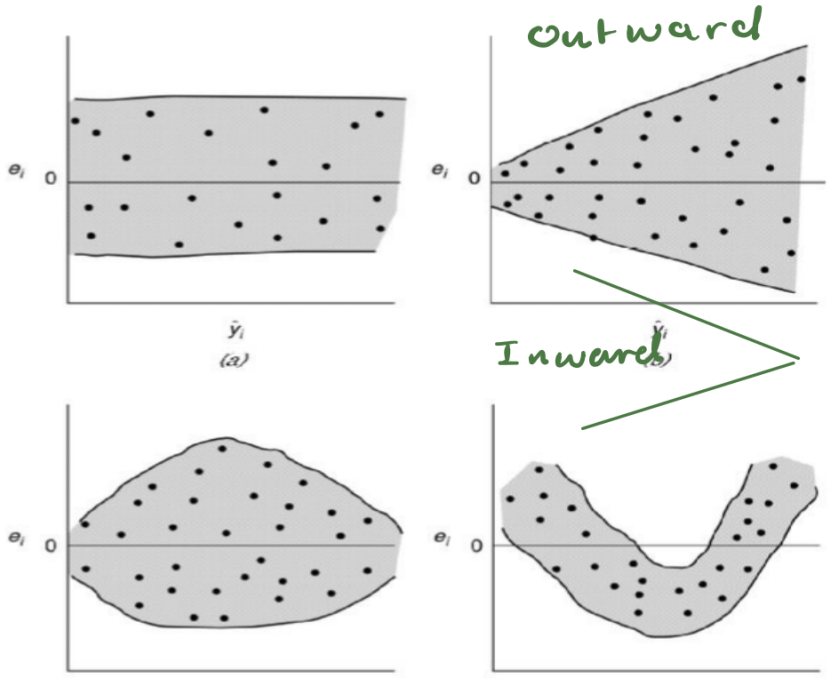


Figure 4.3 Patterns for residual plots: a) satisfactory; b) funnel; c) double bow; d) nonlinear.

a) Good regression model :

scatter in residuals is roughly constant wrt ~~nil~~ $\hat{y}_i$ and centered around $e = 0$ i.e $Var(\varepsilon) = \sigma^2$ is constant variance.

b) Outward Opening funnel : $Var\,\varepsilon \uparrow$ as $\hat{y}_i \uparrow$
Inverted Opening funnel : $Var\,\varepsilon \downarrow$ as $\hat{y}_i \uparrow$

The variance $V(\varepsilon) \neq \sigma^2$ non constant

c) Double Bow Structure :

Indicates a non-constant variance $V(\varepsilon) \neq \sigma^2$ happens when y's proportioned $0 \leq y \leq 1$.

d) Non-Linear Structure :

Other regression variables are needed in the model, consider some extra terms (square term x^2) to the model or apply a transformation to y (log or sqrt).

Why do we plot Res against $\hat{y}_i$ or x but not y?

we assume e_i and y_i are correlated but e_i and $\hat{y}_i$ are not correlated.

Assume there exists a regression model ;

$$e_i = \beta_0 + \beta_1 y_i + \varepsilon_i \qquad \text{SLR}$$

$$LSE \Rightarrow \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{S_{ey}}{S_{yy}} = \frac{\sum (e_i - \bar{e})(y_i - \bar{y})}{\sum (y_i - \bar{y})^2}$$

$$\rho(e, Y) = \frac{S_{ey}}{S_{yy}} = \frac{\sum e_i (y - \bar{y})}{SS_T}, \text{ as } \bar{e} = 0$$

$\bar{e} \sum (y_i - \bar{y}) = 0$
$\bar{y} \sum e_i = 0$

$$= \frac{\sum (e_i - \bar{e})(y_i - \bar{y})}{\sum (y_i - \bar{y})^2} = \frac{\sum e_i y_i}{SS_T}$$

$$= \frac{\sum e_i y_i}{SS_T} = \frac{Y^T e}{SS_T}$$

$$= \frac{Y^T e}{SS_T} = \frac{Y^T (I - H) Y}{SS_T}$$

$(I - H)^2 = (I - H)$ idempotent

$$= \frac{Y^T (I-H)(I-H) Y}{SS_T} = \frac{Y^T (I - H)(I - H) Y}{SS_T}$$

$$= \frac{e^T e}{SS_T} = \frac{e^T e}{SS_T}$$

$$= \frac{\sum e_i^2}{SS_T} = \frac{SS_{Res}}{SS_T}$$

$$= \frac{SS_{Res}}{SS_T} = 1 - \frac{SS_{Reg}}{SS_T}$$

$$= 1 - R^2 = 1 - R^2$$

Assume there exists a regression model ;

$$e_i = \beta_0 + \beta_1 \hat{y}_i + \varepsilon_i \qquad \text{SLR}$$

$$LSE \Rightarrow \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{S_{e\hat{y}}}{S_{\hat{y}\hat{y}}} = \frac{\sum (e_i - \bar{e})(\hat{y}_i - \bar{\hat{y}})}{\sum (\hat{y}_i - \bar{\hat{y}})^2}$$

$$\sum e_i (\hat{y}_i - \bar{y}) = \sum e_i \hat{y}_i - \bar{y} \sum e_i = \frac{\sum e_i (\hat{y}_i - \bar{\hat{y}_i})}{\sum (\hat{y}_i - \bar{\hat{y}})^2}$$

$$= 0 - 0 = \frac{\sum e_i \hat{y}_i}{\sum (\hat{y}_i - \bar{\hat{y}})^2}$$

$$= 0$$

$$\hat{Y}^T e = (X\hat{\beta})^T e = \frac{e^T \hat{Y}}{\sum (\hat{y}_i - \bar{\hat{y}})^2}$$

$$= \hat{\beta}^T X^T e$$

$$= ((X^T X)^{-1} X^T Y)^T X^T e = \frac{Y^T (I - H) H Y}{\sum (\hat{y}_i - \bar{\hat{y}})^2}$$

$$= Y^T X (X^T X)^{-1} X^T e$$

$$= Y^T H (I - H) e = \frac{Y^T (H - H^2) Y}{\sum (\hat{y}_i - \bar{\hat{y}})^2}$$

$$= Y^T (H - H^2) e$$

$$= 0 \qquad = 0 \qquad , \; H = H^2$$

$$\sum e_i \hat{Y}_i = e^T \hat{Y} \qquad e^T = ((I - H) Y)^T = Y^T (I - H)$$

$$\hat{Y} = X\hat{\beta} = X (X^T X)^{-1} X^T Y = HY$$

$$\mathcal{S}(e, \hat{Y}) = 0 \Rightarrow Cov(e, \hat{Y}) = Cov((I-H)Y, X\hat{\beta}) = Cov((I-H)Y, X(X^T X)^{-1} X^T Y)$$

$$= Cov((I-H)Y, HY) = (I-H) H \, Cov(Y, Y)$$

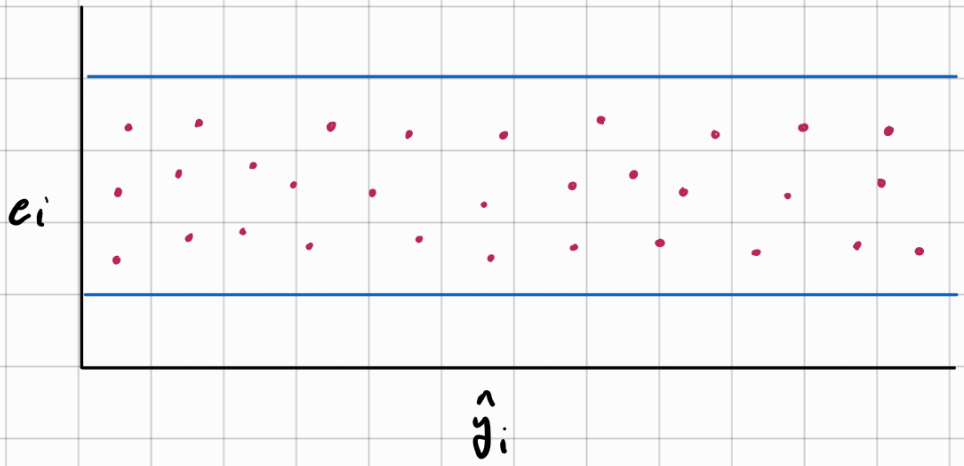
Important: $$= (H^2 - H) \, Var(Y) = (H - H) \sigma^2 = 0$$

$\hat{\beta}_1$ for e_i & y_i : $\hat{\beta}_1 = 1 - R^2$

If $R^2 = 1$ then the slope $\hat{\beta}_1 = 0$, in e_i and y_i plot there is nothing wrong with the model as there is a theoretical justification for the relationship between the residual and y_i, it is very likely that the residuals will not be contained within a horizontal band centered at $e = 0$, there will always be a slope

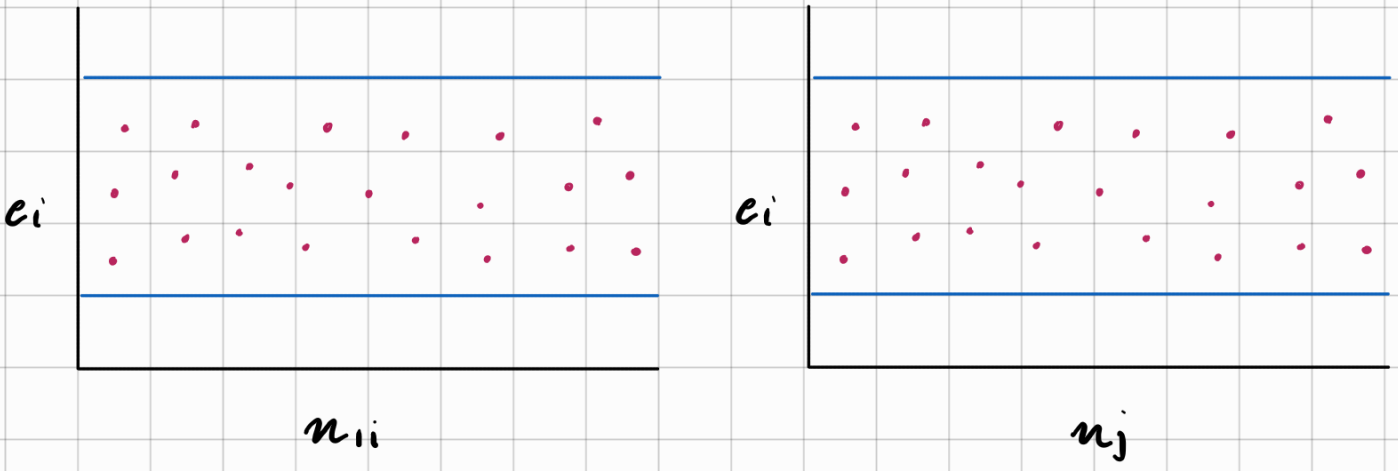
of $(1-R^2)$ when $R^2 \neq 1$.

But when we plot e_i vs $\hat{y}_i$ we can get



there is no linear relationship between e_i and $\hat{y}_i$

3) Partial Regression & partial residual plot



- PRP consider the marginal role of the regressor x_j given other regressors that are already present in the model (MLR case).

- In this plot, the response y_i and regressor x_j are both regressed against the other regressors in the model (except x_j) residuals are obtained from each regression.

$y = f_1(x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k)$ &
$x_j = f_2(x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k)$

Plot of the residuals against each other gives the marginal role of the regressor x_j on y in the presence of $x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$

Example : Consider a MLR with 2 regressors

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

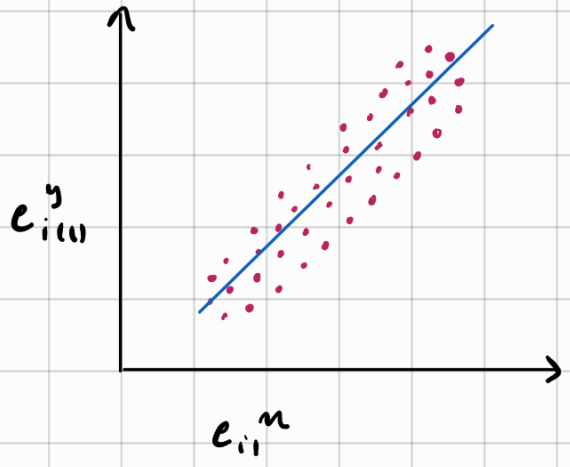
We are interested in the marginal role of x_1 on the response variable Y in presence of the other regressor x_2.

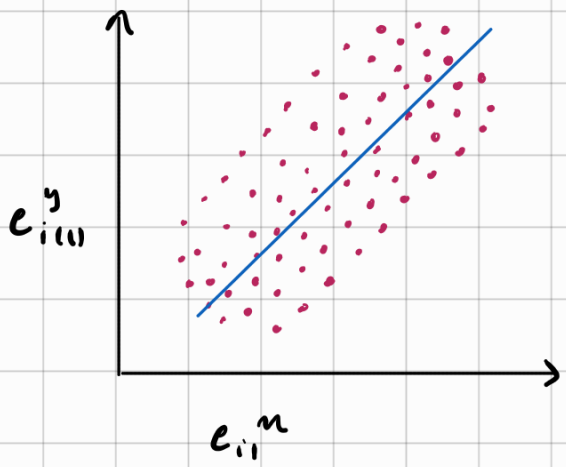
Regression of Y on x_2 : $\hat{Y}_{i(1)} = \hat{\theta}_0 + \hat{\theta}_1 x_{i2}$

Regression of x_1 on x_2 : $\hat{x}_{i(1)} = \hat{\alpha}_0 + \hat{\alpha}_1 x_{i2}$

Residuals : $e^{y}_{i(1)} = Y_i - \hat{Y}_{i(1)}$ eliminates the effect of x_2 from Y

$e^{x}_{i(1)} = x_{i1} - \hat{x}_{i(1)}$ eliminates the effect of x_2 from x_1

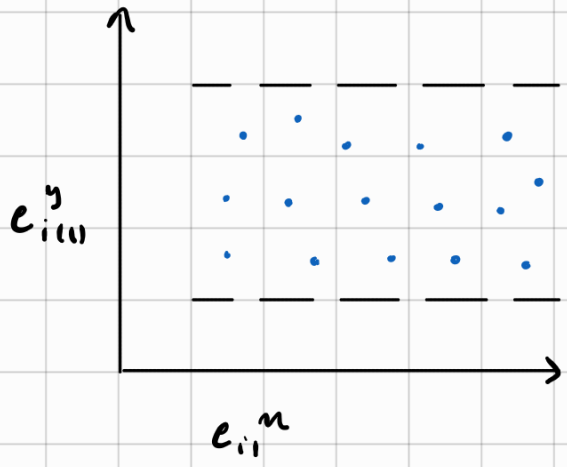




Residual of Y has a linear relationship with the residuals of x_1

Adding x_1 in the model will improve the regression

framework so the marginal effect of x_i is useful as there are points in the data not explained.



In this case we do not suggest including x_i in the data but before eliminating of β_i & do some testing.

General Case: The partial residual of y for n_j as defined as $e^{y}_{i(j)} = y_i - \hat{y}_{i(j)}$

$\hat{y}_{i(j)}$ is the prediction based on K-1 regressors (not n_j)

$\hat{e}_{i(j)}$ represents the variability in y_i that isn't explained by a model that excludes the regressor n_j.

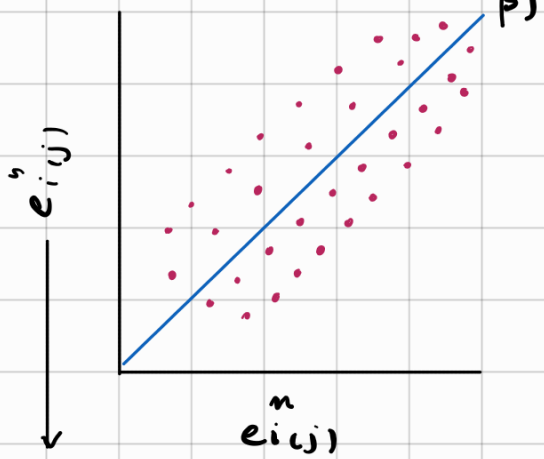
The partial residuals of n_j is defined $\hat{e}_{i(j)} = n_{ij} - \hat{n}_{i(j)}$ where $\hat{n}_{i(j)}$ is the prediction of the regressor value n_{ij} from regression of n_j on all other regressor variables.

$e^{n}_{i(j)}$ is the variation that cannot be explained by (K-1) other regressors.

$$e^{y}_{i(j)} = \beta_j e^{n}_{i(j)} + \varepsilon_i^*$$

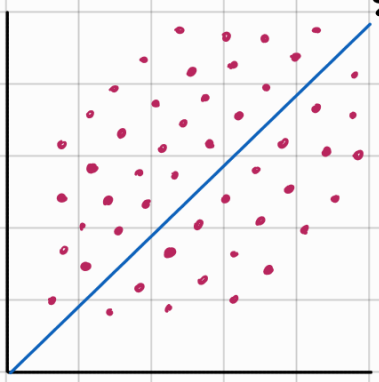
β_j is the slope of the partial residual plots.

β_j slope

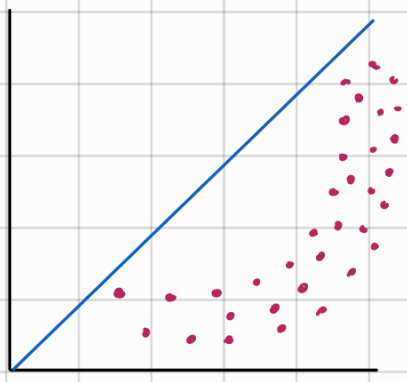


partial residual of y for x_j

less scattered plot indicating a strong relationship between y and x_j



Points are scattered



Curvi linear Bond x_j is not linearly related to y either get a higher term of x_j or transform it ($1/x_j$ or $\log(x_j)$)

* Checking Normality Assumption :

If not linear LSE and all the testings we know do not hold and we have to change our method of estimation

Q - Q Plot : Graphical tool used to assess normality. It plots the sample quantile against the theoretical quantiles on the horizontal quantile where the original data point x_i value is called sample quantile and the expected z-score for the point x_i is called the theo. quantile.

If the data comes from a normal distribution then the theoretical and sample quantiles agree hence giving an approximately straight line Q - Q plot. Otherwise, the plot is not a straight line.

* Statistical Tests of Normality :

1) KS - Test (Kolmogrev Smirnov Test) :

Let $X_1, X_2, \ldots, X_n$ come from a continuous distribution p. We want to test the following hypothesis :

H_0 : Samples come from p
H_1 : Samples don't come from p

Let F be the CDF of X under H_0 then the empirical distribution (observed) function is :

$$F_{obs} = \frac{I(X \leq n)}{\text{Total \# of obs.}}$$

Let F_p be the CDF associated with H_0. Then, the KS - test statistic is given by

$$Dn = \max\{|F_p(n) - F_{obs}(n)|\}$$

Procedure :

1) Order your dataset. Let the ordered data be $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ so the $F_{obs} = \frac{I(X \leq n)}{\text{T. \# of obs}} = \frac{i}{n}$

2) Find F_p for $X_{(i)}$ for each i and calculate the

value of $|F_n(x) - F_{obs}(x)|$ for n values of x and the maximum of these values is D_n.

3) If we use $\alpha = 0.05$ then $D_{c,0.05} = \frac{1.36}{\sqrt{n}}$, H_0 rejected if $D_n > D_{c,0.05}$

2) Shapiro Wilk's Test :

Calculate the W-statistic to check whether $X_1, \ldots, X_n$ comes from a normal distribution (KS applies to any dist but SW only applicable for normal).

The W-statistic is given by : $W = \frac{\sum_{i=1}^{n} (a_i X_{(i)})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$

where $X_{(i)}$ is the ordered sample values and $(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{c}$ for $c = \| V^{-1} m \|$ where V is the co-variance - variance matrix of the ordered sample and $m = (E(X_{(1)}) \ldots E(X_{(n)}))^T$ (mean values of the ordered sample.

If p-value < 0.05, we reject $H_0 :=$ X_i's come from normal against H_1 : X_i's are not normal

* Suppose we detected that the data is not normal then we need a Box - Cox transformation.

If there is evidence of non-normality then the standard remedy is to transform the response

variables using a Box-Cox transformation.

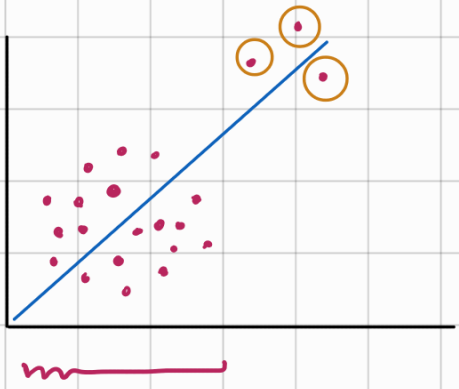
The response variable in Box-Cox method need to be a +ve value and we assume that ∃ a transformation parameter λ (hyper parameter);

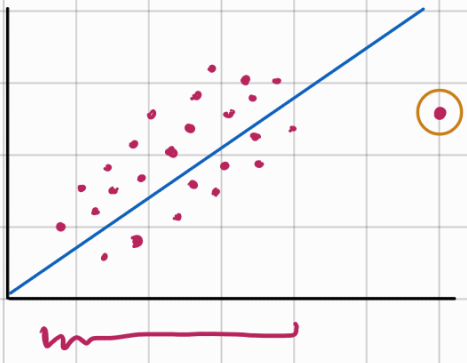
$$Y_I^{(i)} = g(y_i, \lambda) = \begin{cases} \frac{y_i^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \log(y_i), & \lambda = 0 \end{cases}$$

Assume that the transformed response $t(Y) \sim$ multivariate normal distribution, we try to maximize the likelihood function $\mathcal{L}$ of response variable wrt λ.

shift μ
decrease σ^2

* Outlier : Is a data point whose response Y does not follow the general trend of the data





Leverage obs

If +ve x-coord. has unusual value than most of the obs

Influential obs

If it have moderately unusual x and y coord.

1) Testing For Leverage :

$$MLR : \quad Y = X\beta + \varepsilon$$
$$LSE : \quad \hat{\beta} = (X^T X)^{-1} X^T Y$$

$$\text{Fitted Model} : \quad \hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = HY$$

H plays an important role in identifying leverage point h_{ii} is the i^{th} diagonal element of H.

$$h_{ii} = x_i (X_i^T X_i)^{-1} x_i^T \quad ; \quad X_{n \times K} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}$$

is the standerdized distance of i^{th} obs. from the center of the x-coordinate.

We call a point a leverage point if it has an unusual x-coordinate, high h_{ii} value indicates that the i^{th} obs is a leverage point.

$$\bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{Tr(H)}{n} = \frac{Rank(H)}{n} = \frac{K}{n}$$

as H is idempotent

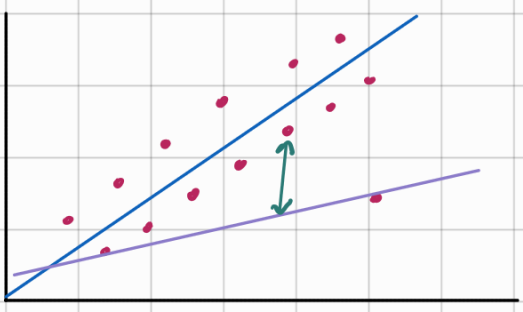
If $h_{ii} > 2\bar{h} = \frac{2K}{n}$ then the i^{th} obs is a possible leverage point

leverage point.

2) Testing for influential point

Influential : If the point influences any part of a regression analysis, the estimate of slope or regression test results.

The removal of an influential data point causes a large change in the fit specially if the data is small.



There are three ways to determine influential obs.

Cook's Statistic :

For i^{th} obs is based on the difference between the predicted response $\hat{Y}$ obtained using the obs. of the predicted response $\hat{Y}_i$ obtained without the i^{th} obs.

$$D_i = \frac{(\hat{Y}_{(i)} - \hat{Y})^T (Y_{(i)} - \hat{Y})}{K \, MS_{Res}} = \frac{\sum_j (\hat{Y}_{(i)j} - \hat{Y}_j)^2}{K \, MS_{Res}}$$

squared euclidean distance between the vector of fitted values & the vector of fitted values when i^{th} the obs is deleted

$$D = (D_1, D_2, \ldots, D_n)$$

If a value D_i is much larger than others then it may indicate that the i^{th} obs are influential.

$D_i > 1$: Indicates a potential high influential observation

* DFFITS : Difference b/w the fit statistics

DFFITS indicates the deletion influence of the i^{th} obs at the fitted value.

For the i^{th} obs., the statistic is given by :

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{MSRes_{(i)}\, h_{ii}}}$$

$\hat{y}_{(i)}$: fitted value obtained without i^{th} obs
$MSRes_{(i)}$: predicted value of MSRes without i^{th} obs

A possible high value influential obs is indicated by :

$$|DFFITS| > 2\sqrt{\frac{K}{n}}$$

* DFBETAS : How much the regression coeff $\hat{\beta}_j$ changes if the i^{th} obs is deleted

$$DFBETAS_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{MSRes_{(i)}\,(X^TX)^{-1}}} \quad \forall_{i=1,\dots,n} \; \forall_{j=0,\dots,K-1}$$

sample coeff.

$\hat{\beta}_{j(i)}$: regression coeff computed without i^{th} obs

A possible high value influential obs is indicated by:

$$|DFBETAS_{ij}| > \frac{2}{\sqrt{n}}$$

$$Y = X\beta + \varepsilon \quad , \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \qquad \begin{aligned} Var(\varepsilon) &= \sigma^2 I \\ Cov(\varepsilon_i, \varepsilon_j) &= 0 \quad \forall i \neq j \end{aligned}$$

When data (y_t, x_t) are collected sequentially in time (time series) the usual assumption of independence of errors is generally not guaranteed such a data is a time series data so $Cov(\varepsilon_i, \varepsilon_j) \neq 0$ so $Cov \neq 0$

* Error autocorrelated / serially correlated correlation b/w errors that are s step apart

$$Cov(\varepsilon_t, \varepsilon_{t+s}) = \rho_s \quad , \quad s = 1, 2, \ldots$$

* Correlation b/w residuals is called lagged correlation & it range between $[-1, 1]$